etami

Bias and fairness in AI: navigating the pitfalls through a use-case in wearables for health

14/06/2023 | Adrian Byrne & Pavlos Sermpezis

Slides available at:

https://github.com/Datalab-AUTH/etami

Bias, discrimination and fairness in AI

14/06/2023 | Adrian Byrne

Slides available at:

https://github.com/Datalab-AUTH/etami

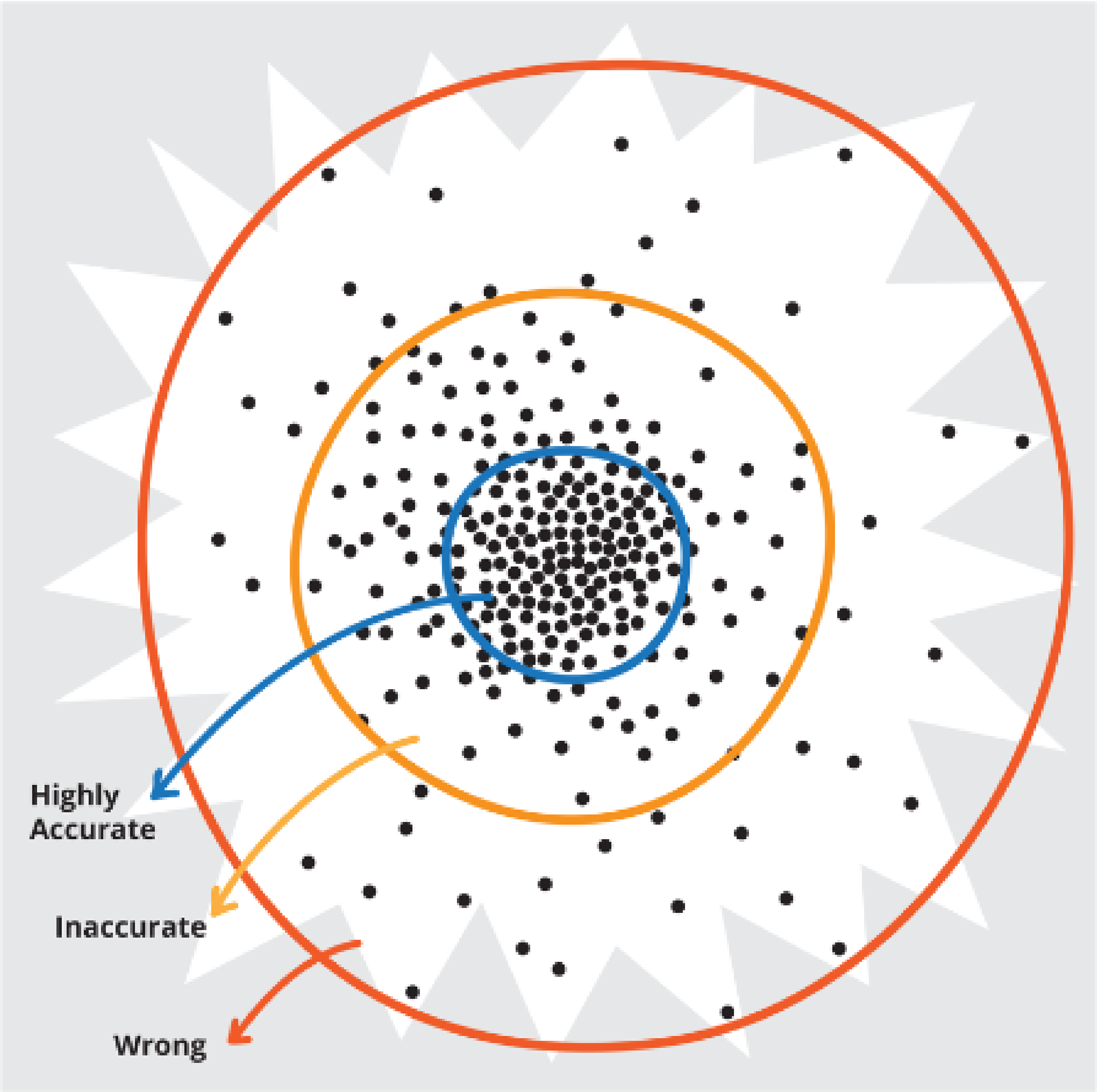
AI is everywhere (almost!)

AI systems can be used in many sensitive environments to make important and life-changing decisions

Therefore, it is crucial to ensure that these decisions do not reflect discriminatory behaviour toward certain groups or populations

The human starburst

Needs of a population plotted in a multi-variate scatterplot showing the accuracy of any statistically determined truth relative to the position within the distribution.



Source: Jutta Treviranus (https://opendatascience.com/collateral-damage-in-the-battle-over-truth/)

Source material

A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

Types of bias: data to algorithm

Measurement bias arises from how we choose, utilise, and measure features

Omitted variable bias occurs when one or more important variables are left out of the model

Representation bias arises from how we sample from a population during data collection process

Aggregation bias arises when false conclusions are drawn about individuals from observing the entire population

Sampling bias arises due to non-random sampling of subgroups

Linking bias arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behaviour of the users

Types of bias: algorithm to user

Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm

User interaction bias is a type of bias that can get triggered from two sources; the user interface and through the user itself by imposing his/her self-selected biased behaviour and interaction

Popularity bias relates to most popular items being exposed more often but popularity metrics are subject to manipulation, e.g. by fake reviews or social bots

Emergent bias occurs as a result of use and interaction with real users. This bias arises as a result of change in population, cultural values or societal knowledge usually some time after the completion of design

Evaluation bias happens during model evaluation via inappropriate/disproportionate benchmarks

Types of bias: user to data

Historical bias is the already existing bias and socio-technical issues in the world and can seep into the data generation process even given perfect sampling and feature selection

Population bias arises when statistics, demographics, representatives and user characteristics are different in the user population of the platform from the original target population

Self-selection bias is a subtype of the selection or sampling bias in which subjects of the research select themselves

Social bias happens when others' actions affect our judgement

Behavioural bias arises from different user behaviour across platforms, contexts, or different datasets

Temporal bias arises from differences in populations and behaviours over time

Content production bias arises from structural, lexical, semantic and syntactic differences in the content generated by users

Discrimination

Bias → unfairness → data collection, sampling, and measurement

Discrimination → unfairness → human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally

Explainable discrimination (justifiable inequality) relates to differences in treatment and outcomes among different groups that can be explained/justified via some attributes in some cases, i.e. legal discrimination

Unexplainable discrimination (unjustifiable inequality) is considered illegal discrimination

Types of discrimination

Direct discrimination happens when protected attributes of individuals explicitly result in non-favourable outcomes toward them

Indirect discrimination is when individuals appear to be treated justly based on seemingly neutral and non-protected attributes. However, protected groups, or individuals, still get to be treated unjustly as a result of implicit effects from their protected attributes

Systemic discrimination refers to policies, customs, or behaviours that are a part of the culture or structure of an organisation that may perpetuate discrimination against certain subgroups of the population

Statistical discrimination is a phenomenon where decision-makers use average group statistics to judge an individual belonging to that group

Fairness

Equalised Odds: A predictor $\hat{Y}$ satisfies equalised odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y. $\mathbf{P(\hat{Y}=1|A=0,Y=y) = P(\hat{Y}=1|A=1,Y=y),\ y \in \{0,1\}}$

Equal Opportunity: A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if $\mathbf{P(\hat{Y}=1|A=0,Y=1) = P(\hat{Y}=1|A=1,Y=1)}$

Demographic/Statistical Parity: A predictor $\hat{Y}$ satisfies demographic parity if $\mathbf{P(\hat{Y}|A=0) = P(\hat{Y}|A=1)}$

Treatment Equality: The ratio of false negatives and false positives is the same for protected group categories

Test Fairness: A score S = S(x) is test fair (well-calibrated) if it reflects the same prediction irrespective of the individual's group membership, R. That is, for all values of s, $\mathbf{P(Y=1|S=s,R=b) = P(Y=1|S=s,R=w)}$

Counterfactual Fairness: Predictor $\hat{Y}$ is counterfactually fair if under any context X=x and A=a, $\mathbf{P(\hat{Y}(U)=y|X=x,A=a) = P(\hat{Y}(U)=y|X=x,A=a')}$

Conditional Demographic/Statistical Parity: For a set of legitimate factors L, predictor $\hat{Y}$ satisfies conditional statistical parity if $\mathbf{P(\hat{Y}|L=1,A=0) = P(\hat{Y}|L=1,A=1)}$

Fairness

Individual fairness: give similar predictions to similar individuals

Group fairness: treat different groups equally

Subgroup fairness intends to obtain the best properties of the group and individual notions of fairness. It picks a group fairness constraint like equalising false positives and asks whether this constraint holds over a large collection of subgroups

Methods for fair machine learning

Pre-processing techniques try to transform the **data** so the underlying discrimination is removed

In-processing techniques try to modify and change learning **algorithms** to remove discrimination during the model training process. In-processing can be used during the training of a model either by incorporating changes into the objective function or imposing a constraint

Post-processing can be used in which the **labels** assigned by the black-box model initially get reassigned based on a function during the post-processing phase

Fairness in Health/Wearables use-cases

14/06/2023 | Pavlos Sermpezis

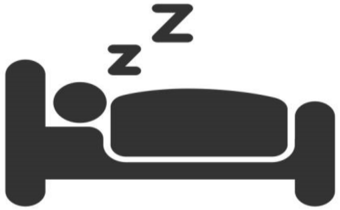
Slides available at:

https://github.com/Datalab-AUTH/etami

Wearables & Ubiquitous Computing

We are all **ubiquitous connected** through our …

… that can monitor our health and physical activity …



Examples of Fairness in health use-cases

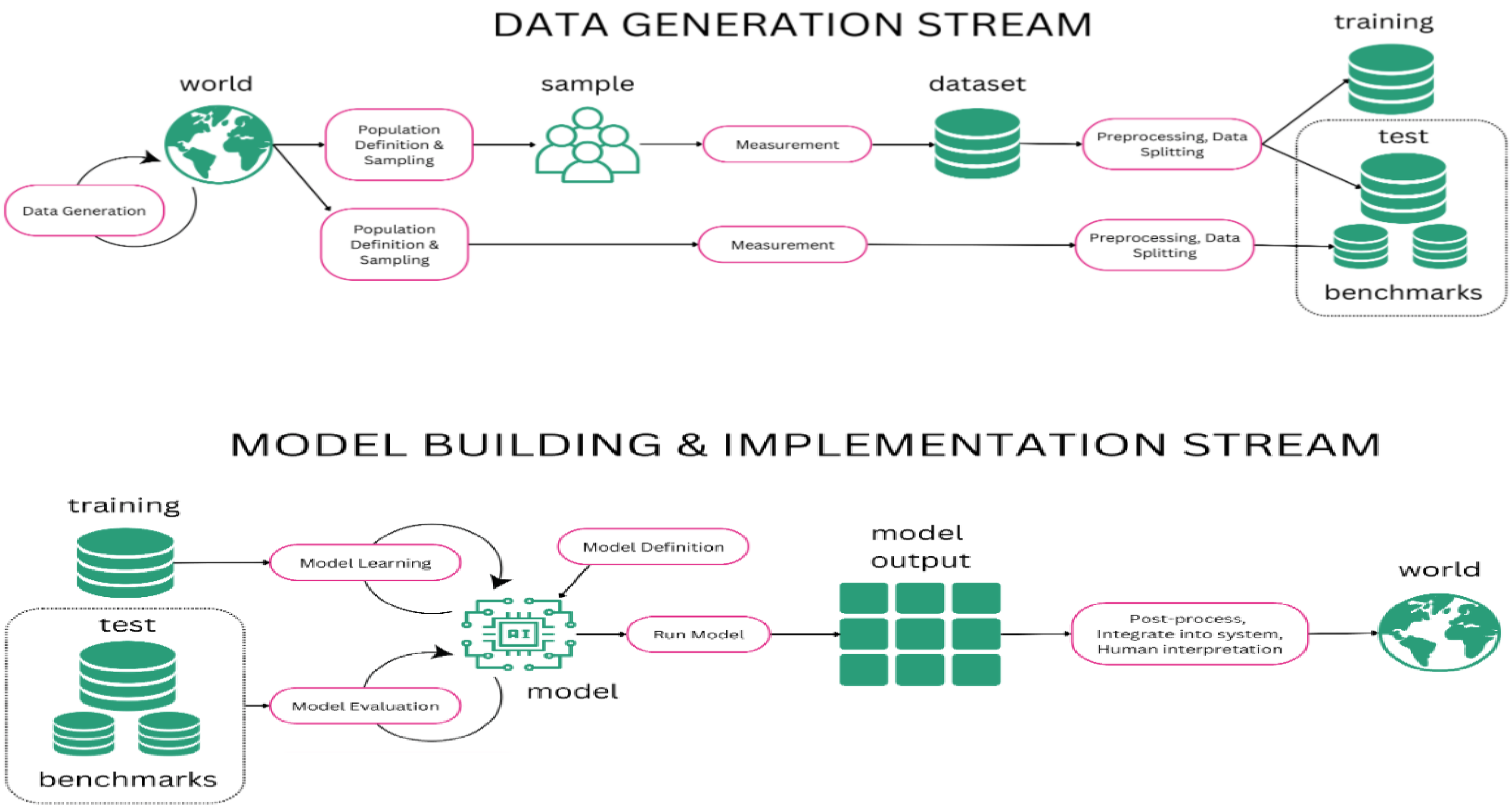
- 🤔 Black patients assigned the same level of risk (i.e., need for extra care) by the algorithm are sicker than White patients; due to using health costs as a proxy for health needs (money spent ~ health need)

- 🤔 Breast cancer recommendations had different accuracy in different age groups; due to quantity of data from previous cases

- 🤔 Medical devices (pulse oximeters) caused delayed treatment for darker-skinned patients during the Covid-19 pandemic due to overestimation of blood oxygen levels in minorities

- 🤔 Female patients are disproportionately misdiagnosed for heart disease (heart attacks), and receive insufficient or incorrect treatment

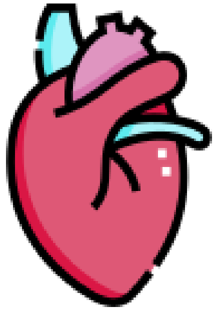
“Sources of bias” in Machine Learning Lifecycle



Reference: Yfantidou, Sofia, et al. "Uncovering Bias in Personal Informatics." *arXiv preprint arXiv:2303.15592* (2023).

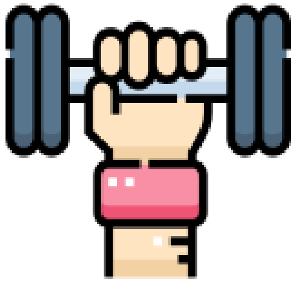
Mobile and Wearable Datasets for Health

MyHeart Counts



Cardiovascular Health
&
Physical Activity

LifeSnaps



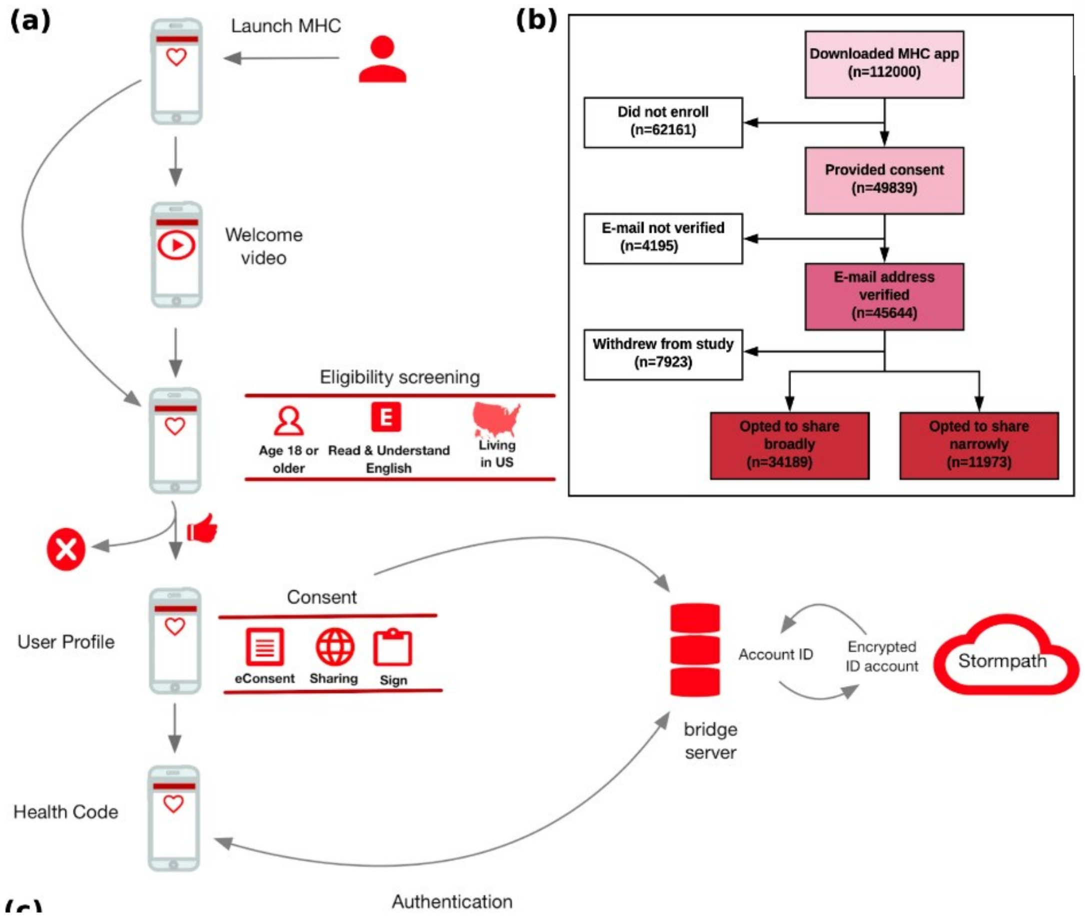
Physical activity
&
Well-being

MIMIC-III

ICU
Health Records

MyHeart Counts

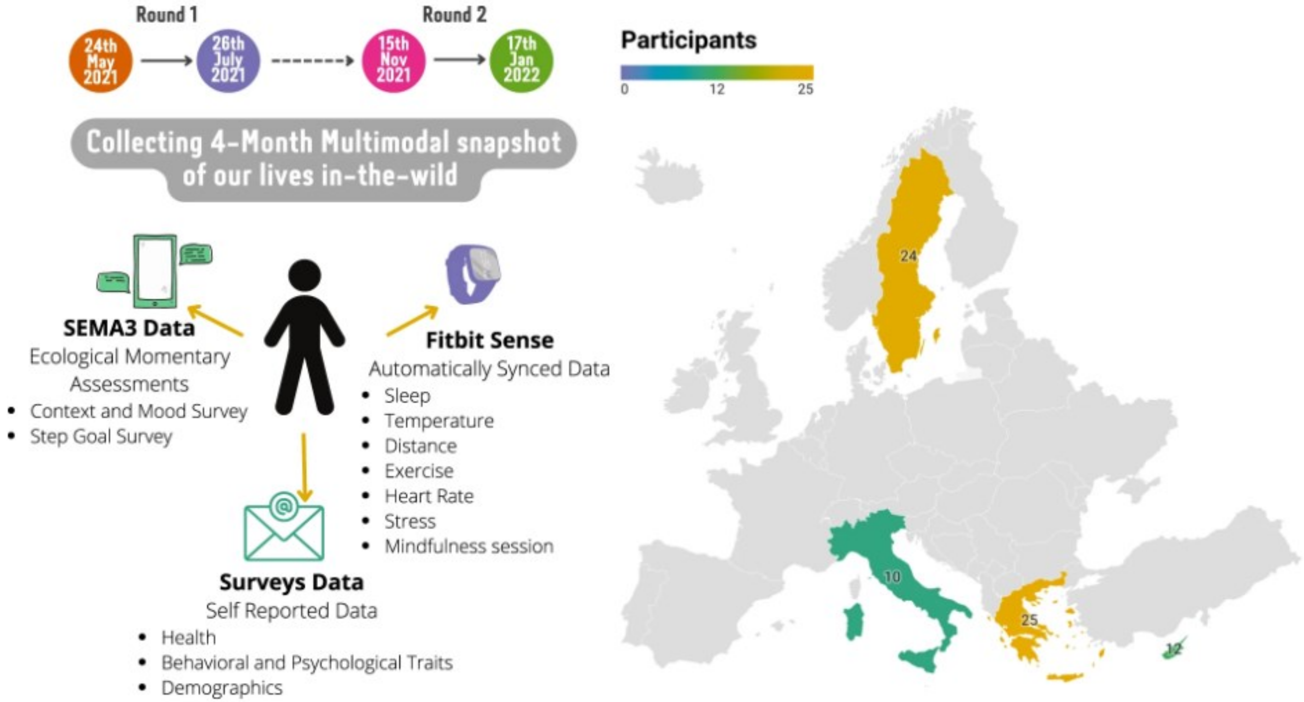
- A **smartphone-based** study of **cardiovascular health**
- Data includes daily **physical activity** and **sleep**, **health questionnaires**, and a **6-minute walk fitness test**
- **Large-scale:** 50000 downloads, of which ~**5000** with activity data
- **Potential use cases:**
  - Activity and sleep prediction to increase user engagement
  - Infer cardiovascular health
  - Recommendation systems



Reference: Hershman, Steven G., et al. "Physical activity, sleep and cardiovascular health data for 50,000 individuals from the MyHeart Counts Study." Scientific data 6.1 (2019): 24.

LifeSnaps

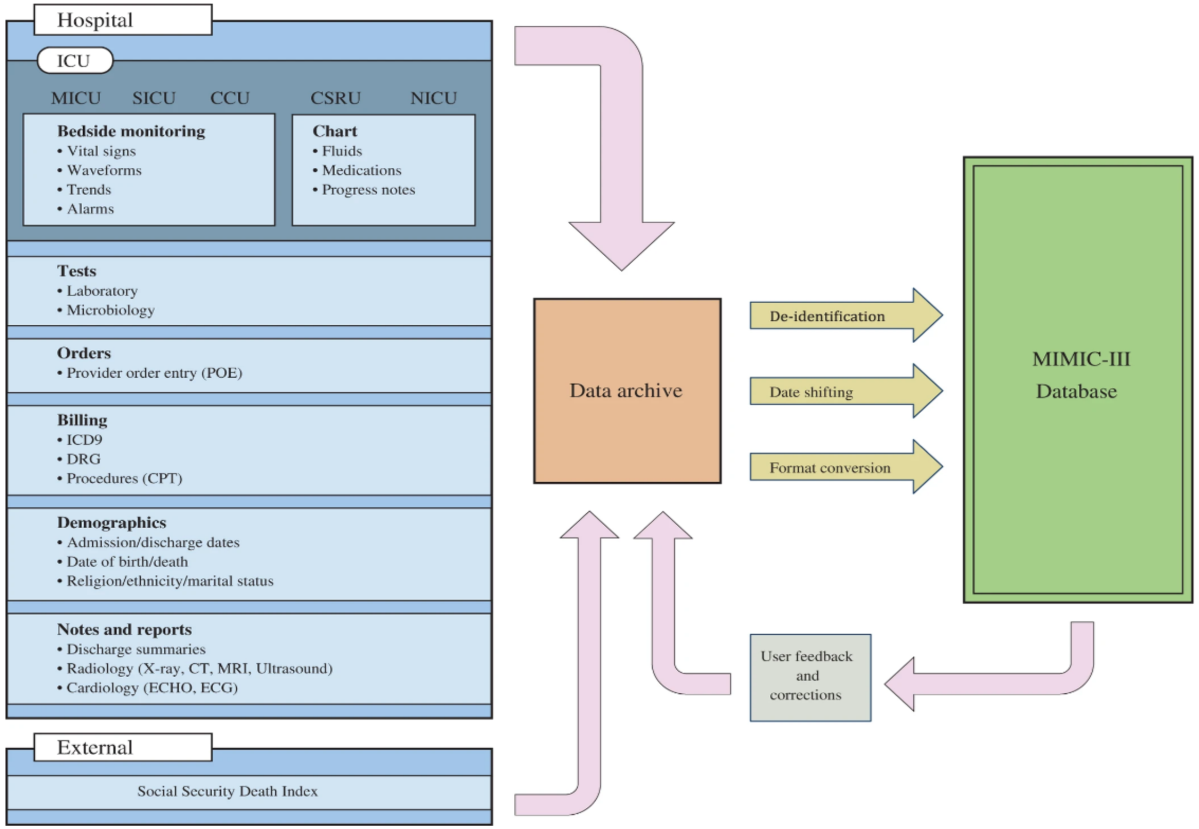
- Studies the **association** between **physical activity patterns, sleep, stress, and overall health, behavioral traits and psychological states**
- Data includes daily physical activity and sleep, psychological and mental health questionnaires, and Ecological Momentary Assessment (EMA) responses
- **Medium-scale:** 71M rows from **71 users** and **35 data modalities**
- **Potential use cases:**
  - Predict physiological data to increase user engagement
  - Infer personality and psychological traits



Reference: Yfantidou, Sofia, et al. "LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild." *Scientific Data* 9.1 (2022): 663.

MIMIC-III

- Data relating to **patients admitted to critical care units** at a large tertiary care hospital in the US
- Data includes **vital signs**, **medications**, notes by care providers, **diagnostic codes**, imaging reports, etc.
- **Large-scale:** 42276 ICU stays of **33798** unique **patients**
- **Potential use cases:**
  - Predict length of stay
  - Predict decompensation
  - Predict mortality



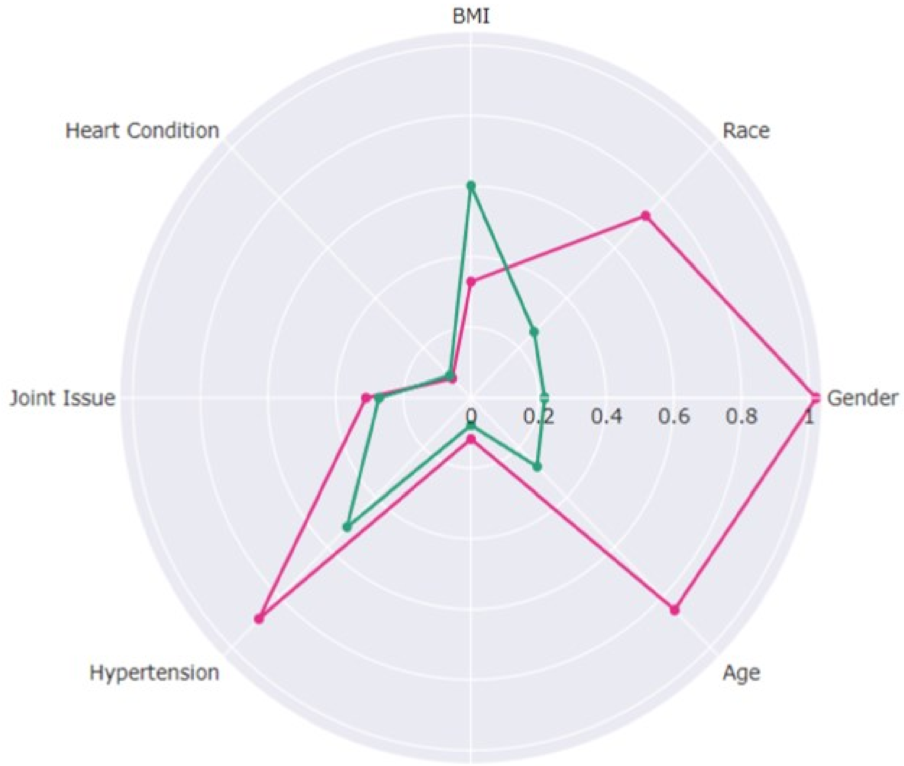
Reference: Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3.1 (2016): 1-9.

Data Biases (incl. Historical, Representation)

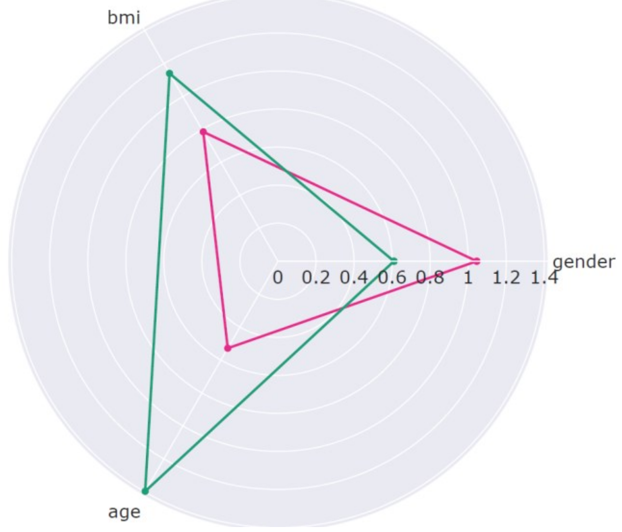
Dataset might not represent the **real target population.**

Real vs. Dataset Population Ratio Comparison (#Minority / #Majority Class Instances)

MyHeart Counts



LifeSnaps



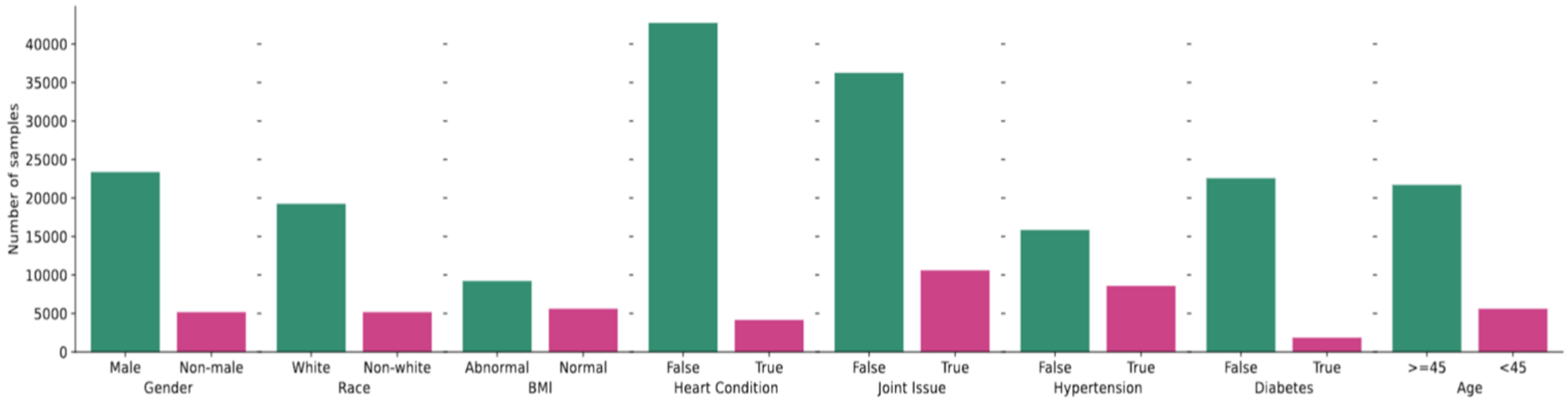
MIMIC-III

Real Ratios
Dataset Ratios

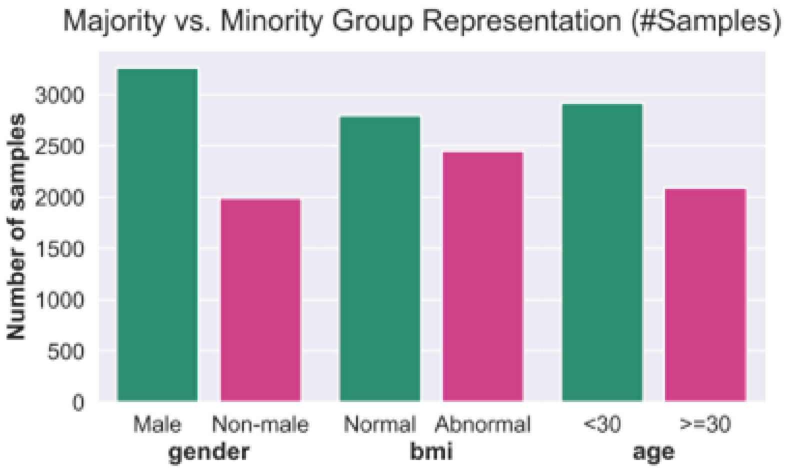
Data Biases (incl. Historical, Representation)

Dataset might include **underrepresented groups** even if sampled perfectly.

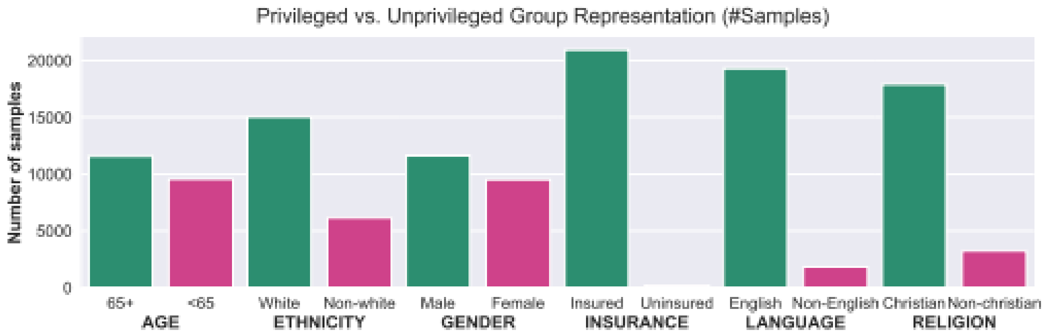
MyHeart Counts



LifeSnaps



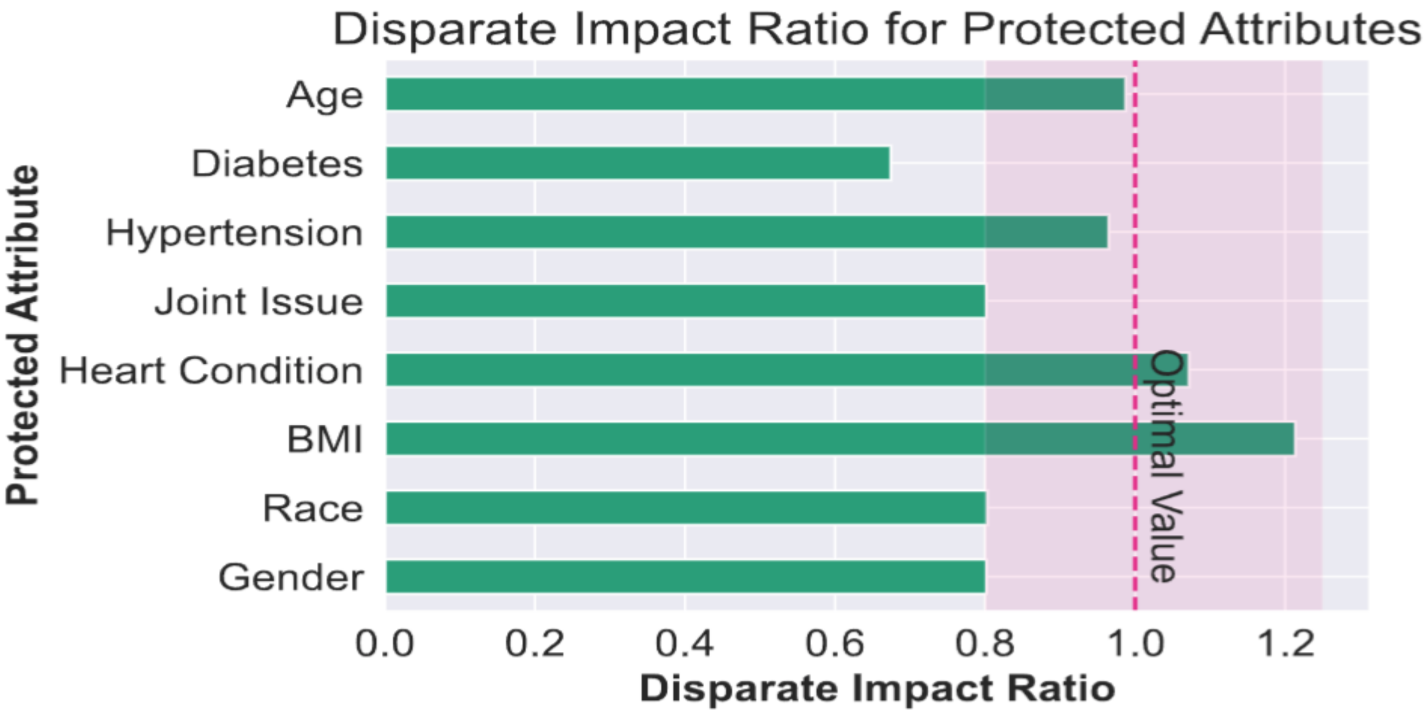
MIMIC-III



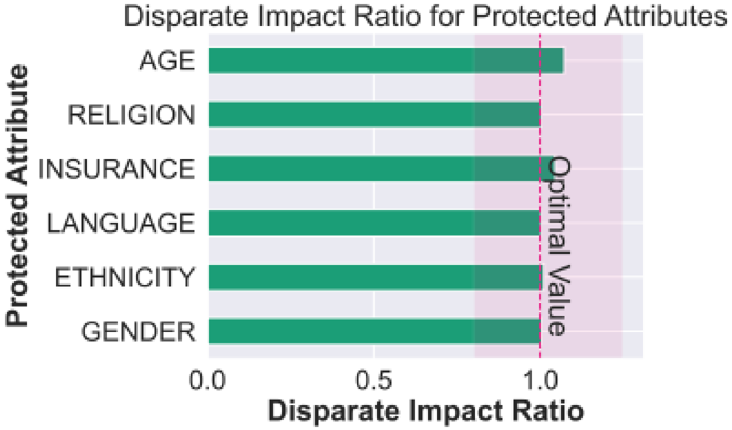
Data Biases (incl. Historical, Representation)

Dataset might suffer from **limited or uneven** sampling method.

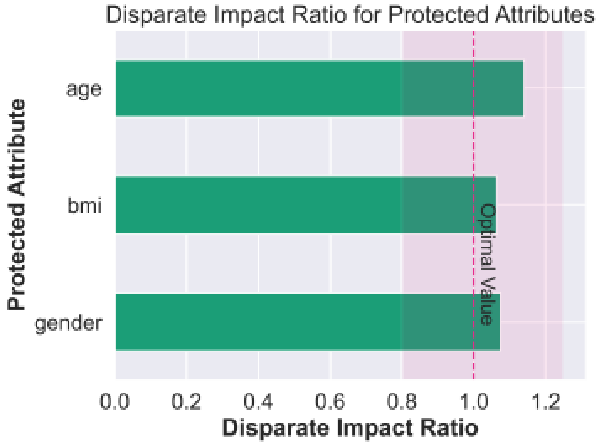
MyHeart Counts



MIMIC-III



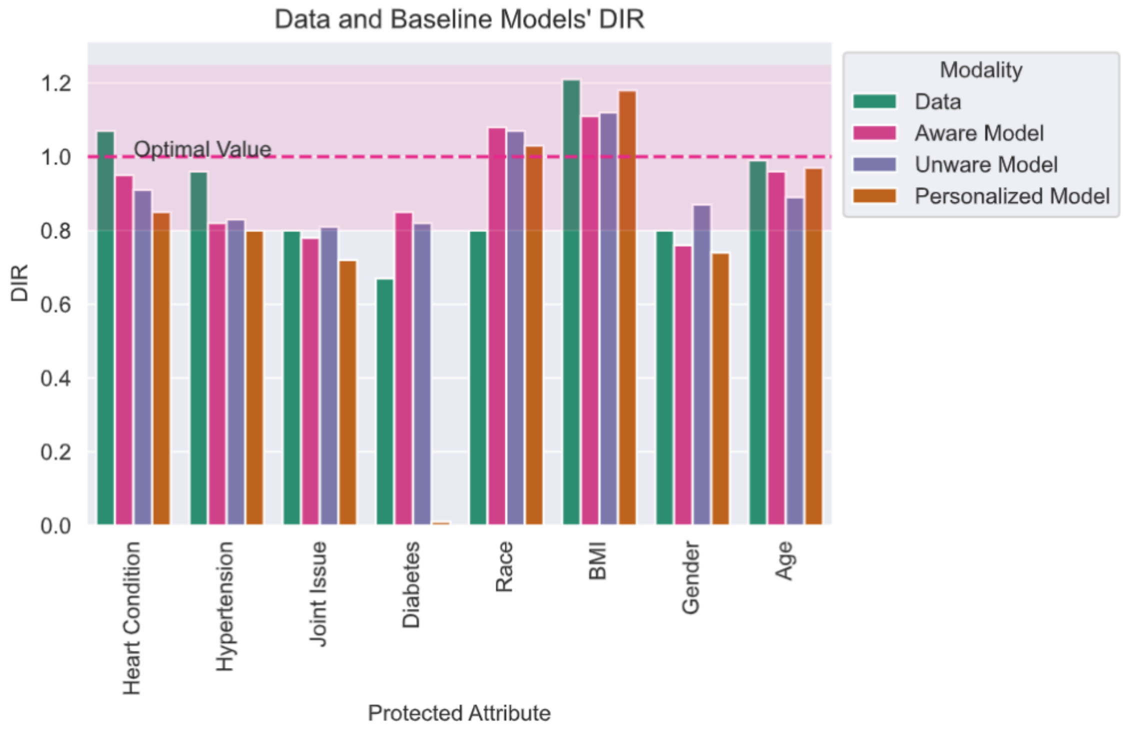
LifeSnaps



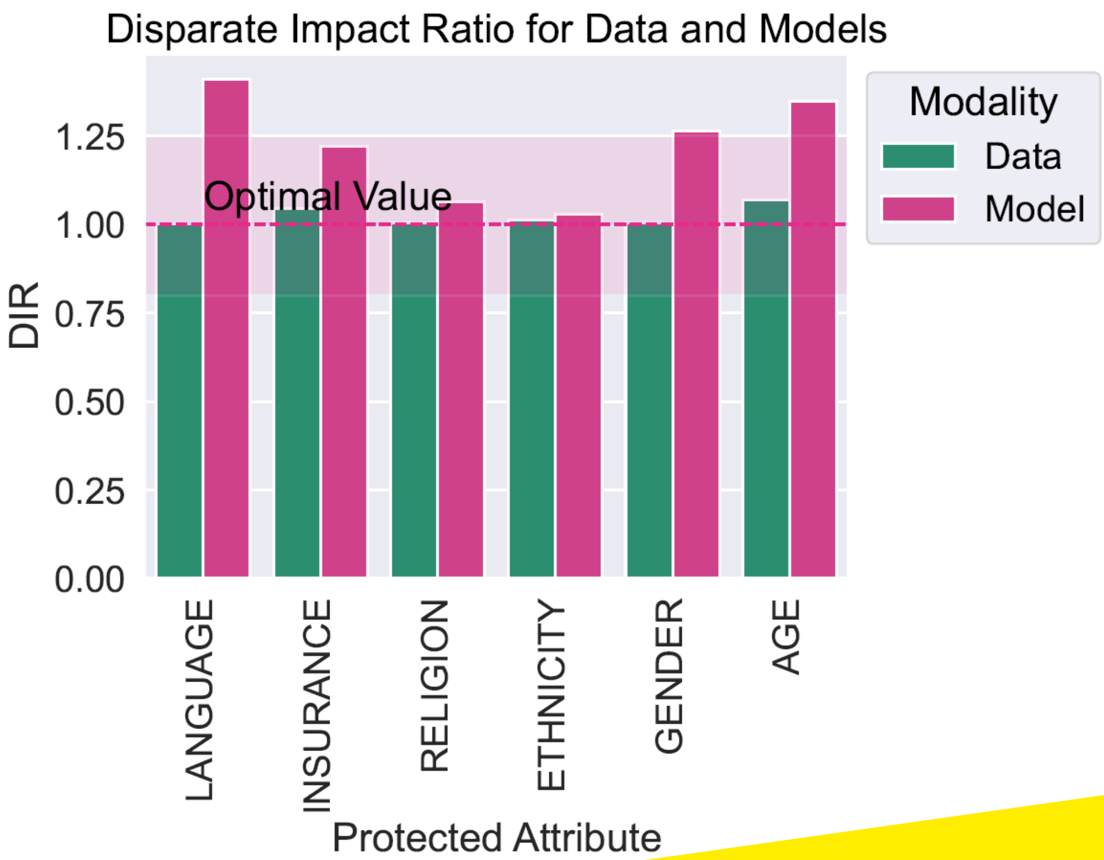
Model Biases (incl. Learning, Aggregation & Evaluation)

Models amplify biases

MyHeart Counts



MIMIC-III



- Aware models **propagate** (joint issue, diabetes, gender) **or amplify** (hypertension) data biases
- Unaware models are **not foolproof** against data bias

A concluding precursor

14/06/2023 | Adrian Byrne

Slides available at:

https://github.com/Datalab-AUTH/etami

Source material

WHY FAIRNESS CANNOT BE AUTOMATED: BRIDGING THE GAP BETWEEN EU NON-DISCRIMINATION LAW AND AI

Sandra Wachter,[1] Brent Mittelstadt,[2] & Chris Russell[3]

The challenge

There exists a substantial literature concerning bias, discrimination and fairness in AI and machine learning

Connecting this work to legal non-discrimination frameworks is essential to create tools and methods that are practically useful across divergent legal regimes

$$A = \frac{Number\ of\ selected\ people\ in\ advantaged\ group}{Total\ number\ of\ people\ in\ advantaged\ group}$$

$$D = \frac{Number\ of\ selected\ people\ in\ disadvantaged\ group}{Total\ number\ of\ people\ in\ disadvantaged\ group}$$

$$\boldsymbol{Demographic\ (dis)parity = D > A}$$

$$A_{\boldsymbol{R}} = \frac{Number\ of\ selected\ people\ in\ advantaged\ group\ \boldsymbol{with\ attributes\ R}}{Total\ number\ of\ people\ in\ advantaged\ group\ \boldsymbol{with\ attributes\ R}}$$

$$D_{\boldsymbol{R}} = \frac{Number\ of\ selected\ people\ in\ disadvantaged\ group\ \boldsymbol{with\ attributes\ R}}{Total\ number\ of\ people\ in\ disadvantaged\ group\ \boldsymbol{with\ attributes\ R}}$$

$$\boldsymbol{Conditional\ demographic\ (dis)parity = D_R > A_R}$$

A proposal

The authors propose summary statistics based on conditional demographic (dis)parity (CDD) as the cornerstone of a coherent strategy to ensure procedural regularity in the identification and assessment of potential discrimination caused by AI and automated decision-making systems

The measure respects **contextual equality** in EU non-discrimination law by not interfering with the capacity of judges to contextually interpret comparative elements and discriminatory thresholds on a case-by-case basis

Thank you!

Have any questions?

14/06/2023 | Adrian Byrne & Pavlos Sermpezis

Slides available at:

https://github.com/Datalab-AUTH/etami